ARPA ORDER NO.: 189-1
3D20 Human Resources

R-1348-ARPA
December 1973

The Role of Experimentation in Manpower Planning

Gus W. Haggstrom

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

25th Year
Rand
SANTA MONICA, CA. 90406

The research described in this Report was sponsored by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

ARPA ORDER NO.: 189-1
3D20 Human Resources

R-1348-ARPA
December 1973

The Role of Experimentation in Manpower Planning

Gus W. Haggstrom

A Report prepared for

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

Rand
SANTA MONICA, CA. 90406

Published by The Rand Corporation

PREFACE

This report is part of Rand's DoD Training and Manpower Management Program, sponsored by The Human Resources Research Office of the Defense Advanced Research Projects Agency (ARPA). The purpose of this research program is to bring new methodologies to bear on present and future military manpower problems. This report discusses the role of controlled experimentation in the field of personnel management and lists the advantages and shortcomings of this methodology relative to other information-gathering techniques.

As the military undergoes the transition to an all-volunteer force, there is an urgent need for more controlled experimentation to assess the effectiveness of proposed changes in personnel practices and to provide better knowledge of the military's utilization of human resources. The military services have already taken steps to expand their personnel reporting systems and data banks. New computer-based models and techniques have become more widely used for tracking the flow of personnel and for evaluating the potential effect of altering one or more of the basic conditions of service. But these efforts should be supplemented by well-conceived pilot studies to reevaluate some of the key parameters affecting manpower supply and utilization in the new environment and to provide real tests of the usefulness of changes in personnel practices before they are put into practice.

SUMMARY

Military policymakers will soon be faced with some hard decisions in adjusting to the realities of an all-volunteer force. Recruiting shortfalls, changes in the educational level and attitudes of recruits, and low retention rates in critical occupational specialties will force some major changes. With more and more Congressional pressure on the military budget as an after-effect of rapidly rising personnel costs over the past five years and changes in public attitudes toward the military establishment, the military can no longer afford the luxury of implementing costly, unproved recruitment and retention strategies. They must take steps to use their human resources more effectively.

This report recommends that the military make greater use of controlled experimentation in personnel management as an information-gathering technique for evaluating some of the proposed changes in personnel policies and for acquiring reliable data as a basis for decisionmaking and model building. An example of the type of experiment recommended in this report is a pilot study now under way in the Air Reserve Forces to test whether reducing the term of enlistment would have a substantial effect upon recruiting. The experiment consists of permitting enlistment options at a small number of carefully selected locations for a period of several months. In a more comprehensive study of the same type, one could test several types of recruiting incentives simultaneously as well as different levels of recruiting efforts and advertising.

Some commonly used alternatives to pilot studies for gathering information and estimating the potential effects of new programs are: (1) soliciting expert opinion and anecdotal evidence, (2) sample surveys, (3) analyses of nonexperimental data, and (4) computer simulation. The advantages and shortcomings of each of these methods are discussed in Section V.

The principal advantage of pilot studies over these alternatives is that the alternatives depend upon indirect evaluations, but in a pilot study the new programs are put into practice on a small scale. Ordinarily, a control group is selected to establish a basis for making more reliable inferences, and efforts are made to delimit the effects of extraneous factors upon the process under study. Seeing how the new policies work in practice and having more reliable measures of performance to act upon, the policymakers can compare the options more precisely, eliminate those policies that do not work, and correct the minor flaws in those that do. This sequential process not only results in better programs but facilitates the implementation

of the full-scale programs and minimizes the negative effects of those undesirable programs that looked good on paper.

In addition to permitting finer comparisons of various options, the process of designing the experiment, carrying out the pilot study, and analyzing the data usually results in other benefits: increased knowledge of the system and how it works, incidental discoveries of valuable information, better quality control of the process under study, and the isolation of unusual experimental subjects. These and other benefits of controlled experiments are discussed in Sections III and IV. The shortcomings of pilot studies are listed in Section VI. A discussion of some considerations involved in designing a pilot study is provided in Section VII.

The last section discusses the military's obligation to experiment. Since the military services are involved in the distribution and utilization of scarce resources —manpower, equipment, time, and money—and the losses associated with poor personnel programs can be extremely costly, the military has a special obligation to strive for more efficient and effective personnel management methods. Testing more of their personnel programs through controlled experiments is a step in that direction.

ACKNOWLEDGMENTS

The author wishes to express a special debt of gratitute to Bernard Rostker for undertaking a leading role in promoting controlled experimentation as an important aspect of military manpower planning and for inviting the author to join him in the design and analysis of experiments that are now underway in the Air Force. The author also wishes to thank Richard V. L. Cooper for his role in expanding Rand's participation in experimentation to the other services and for stimulating the author to write up his thoughts on the benefits and hazards of using experimentation to test personnel policies. Finally, the author wishes to thank William Jones, Robert Klitgaard, Carl Morris, Joseph Newhouse, and Daniel Relles for their many helpful comments on an earlier draft of this report.

CONTENTS

I. INTRODUCTION

During the period of transition to an all-volunteer force, there is an urgent need for controlled experimentation in the military to assess the effectiveness of current and proposed personnel programs, especially those associated with manpower procurement and utilization. In many cases in the past, opportunities for evaluating new programs have been lost through the servicewide implementation of new practices that had neither pilot studies before nor follow-up studies after implementation to analyze their effectiveness. As a result, the military have accumulated a varied assortment of personnel programs of unknown worth, and they have gained little reliable information to guide the choice of future policies. This report discusses the need for controlled experimentation in various aspects of personnel management and indicates the advantages and shortcomings of this type of experimentation relative to other techniques for evaluating policies and gathering information.

Although this study is primarily addressed to personnel management administrators, the arguments for increasing the amount of controlled experimentation in the military apply equally well to other areas, such as systems acquisition, research and development management, and billet construction. However, the major challenges facing the military today in establishing an all-volunteer force are in the field of personnel management. There has been much less use of controlled experimentation in this field in the past than in, say, weapons-related research. In the near future, expenditures for experimentation in the personnel field would surely pay great dividends as the military explore the alternative uses of the personnel budget to increase recruitment and retention and at the same time examine ways to alter the composition and use of the force in response to changing conditions.

II. THE PROBLEMS

Policymakers in the military are now confronted with a number of pressing personnel problems due to the elimination of the draft, the reduction and restructuring of the military forces, and changes in American society. The military services face recruiting shortfalls, changes in the educational level and attitudes of enlistees, low retention rates in critical occupational specialties, and an almost endless list of other personnel problems. Having very limited information on the relative merits of the alternative options open to them, policymakers are being forced to make piecemeal adjustments to current policies without fully understanding the consequences.

A number of critical decisions must be made soon to adjust to the realities of an all-volunteer force. Perhaps the most pressing problems are in the area of procurement, where the costs associated with wrong decisions can be astronomically high. The recruiting requirement for the active forces for fiscal year 1974 is estimated at over 350,000 enlisted men.[1] Even ignoring the additional problems of recruiting officers and reservists for the moment, one can see that meeting this requirement alone, year after year, means recruiting one out of every six men of the appropriate age group into the armed forces. In the absence of further recruiting incentives, sizable shortfalls are projected.

Meanwhile, Congressmen are becoming more and more concerned about rising personnel costs in the military. Pay and related costs for military personnel and civilians in the Department of Defense have risen from $32.6 billion in FY 1968 to $41.8 billion in FY 1973, whereas the size of the active forces dropped from 3.5 to 2.2 million during the same five-year period. As a percentage of total DoD expenditures, pay and related costs increased from 42 percent in FY 1968 to 56 percent in FY 1973.[2]

To meet the recruiting requirements, military policymakers have many options open to them. But most of the options carry high price tags, some written in invisible ink, and the relative advantages and shortcomings of the various options are for the most part unknown. Operating within current budgetary constraints and expecting tighter and tighter budgets in the near future, the military cannot afford large-scale uncertain lunges that lead to exorbitant procurement strategies. Costly incentives

[1] The Comptroller General of the United States, *Report to the Congress: Problems in Meeting Military Manpower Needs in the All-volunteer Force*, General Accounting Office, Washington, D.C., May 2, 1973, p. 12.

[2] Ibid., p. 63.

and enlistment commitments that seem promising on paper may have only limited short-term effectiveness. Yet, once these incentives are introduced, they are likely to persist for many years. The recruitment strategies must be shown to be cost-effective over the long run *before* they are implemented servicewide.

What combinations of recruiting incentives now under consideration would yield the best results for each dollar expended? How can the high-skill occupational specialties be manned using currently available policy options? Lateral entry? Civilianization? Should the mental and physical qualifications be lowered or the initial term of enlistment shortened? Should the military attempt to attract more women (or minority members or college graduates)? If so, how? How effective are recruiters? What personal characteristics do the more effective recruiters have and what methods do they use? What types of advertising are effective, and how much? Merely listing some of the questions being asked about recruiting strategies today indicates the enormity and complexity of the manpower problems to be solved in this area. Clearly, there is an urgent need for reliable information on the cost-effectiveness of the various options implicit in the questions above.

One can compile similar lists of questions for the other areas of personnel management in which significant adjustments must be made to cope with current realities: drug abuse, racial tensions, relations between enlisted men and officers, classification problems, changing roles for women. In making these adjustments, the Defense Department would benefit substantially from a greatly expanded research program in the field of personnel management to provide the relevant data bases and assessments of alternative proposals that lead to determination of cost-effective programs. The Defense Department has neglected this field in the past. As the Ginzberg Report, which decried the small amount of research by the military in the human resources area, pointed out, "The DoD spends approximately $1.00 on manpower research for each $1,000 of payroll in contrast to $1.00 for weapons-related research for every $3.00 spent for hardware procurement."[3]

Partial answers to many of the questions confronted by the military today can be obtained through sample surveys, computer simulation, and analyses using nonexperimental data. These techniques have been used often in the past, but, as will be discussed in a later section, they have decided shortcomings in determining the effectiveness of proposed programs. In many cases the only way to see how well a proposed program will work is to try it out in a carefully controlled experiment. As the prominent statistician G.E.P. Box wrote, "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)."[4]

[3] Task Force on Manpower Research, *Manpower Research and Management in Large Organizations*, Office of the Director of Defense Research and Engineering, Washington, D.C., June 1971 (The Ginzberg Report).

[4] George E. P. Box, "Use and Abuse of Regression," *Technometrics*, Vol. 8, No. 4, November 1966, pp. 625-629.

III. THE BENEFITS OF CONTROLLED EXPERIMENTS

In a "controlled" experiment, the options (or "treatments") under consideration are tested at two or more levels and other factors are held constant insofar as possible to obtain more precise estimates and valid comparisons of the effects of these factors on the responses of interest. For the applications to personnel management considered in this study, the type of experiment suggested here is often referred to as a "pilot study," in which two or more alternative schemes are tested by putting them into practice, using adequate safeguards to assure the validity of the results. Usually such experiments are characterized by the use of randomly chosen "treatment groups" and a "control group" of persons (or units or locations) that do not receive the treatments under study but are also observed during the course of the experiment. The purpose of the control group is to provide a basis for making statistical comparisons with the treatment groups.

An experiment might have only a single treatment group and a control group. Suppose, for example, one wanted to test the effect of a particular type of advertising campaign on the number of first enlistments into a service or reserve component. Suppose further that there are a number of relatively homogeneous units (or locations) that could be used in the study. Then the experiment could be carried out by first choosing a certain subset of the units at random and subjecting them to the advertising campaign. Here the control group would consist of the units that were not chosen to receive the advertising. After observing the numbers of enlistments in both groups over a period of time, one could assess the effectiveness of the advertising by comparing the average number of reenlistments for the two groups. If instead of testing a single level of advertising expenditure, one wished to test two or more levels, this could be done by setting up a treatment group for each of the levels.

A well-designed pilot study can test the relative effectiveness of several alternative proposals simultaneously. For example, to assess the efficacy of various types of recruiting efforts, one might consider experimenting with: (a) different types of advertising, (b) different levels of use of recruiters, and (c) various recruiting incentives. Testing these factors simultaneously would also provide valuable information about the "interactions" among the various components of a recruiting effort. The use of advertising in units having recruiters may be far more beneficial than the combined contributions of just advertising alone or just recruiters alone.

In addition to providing more reliable information than one can usually acquire through alternative methods, a pilot study often uncovers some negative side effects

that might not be foreseen before experimentation. For example, in testing their highly successful Experimental Volunteer Army Training Program (EVATP), investigators from the Human Resources Research Organization (HumRRO) discovered that there were major difficulties associated with the institutional changes required to convert from conventional training to EVATP, in part because of resistance from the instructors.[5] As a hypothetical example, an experiment to test the effectiveness of a reduction in the reservists' enlistment tour length from six years to three may reveal that, although the number of enlistees would surely increase with a shorter term of enlistment, the consequent loss in man-years would be detrimental to the reserves. Perhaps an enlistment bonus to attract servicemen into one occupational specialty would merely take recruits away from other specialties with no net gain in accessions and create bitterness and low morale among the other enlisted men at the same time.

Often the effectiveness of a new policy is measured not by a single response but by a number of criteria. In evaluating a new method of training, for example, one would surely be interested in other measures than just the individual test scores to determine how well each trainee learned the course material. Training time, instructional costs, attrition rates, morale factors, and administrative ease are other factors that might be evaluated directly or indirectly during the course of the experiment. It is unlikely that a radically new training process would work well in all respects. One of the important attributes of pilot studies as compared with across-the-board implementation is that one can eliminate those programs with serious flaws from further consideration as well as any operational "bugs" in the desirable programs, and thereby enhance the chances for success.

[5] J. E. Taylor, E. R. Michaels, and M. F. Brennan, *The Concepts of Performance-Oriented Instruction Used in Developing the Experimental Volunteer Army Training Program*, Human Resources Research Organization, Technical Report No. 72-7, Alexandria, Virginia, March 1972.

IV. INCIDENTAL FINDINGS AND SERENDIPITY

I have emphasized the need for controlled experimentation to uncover the unforeseen negative aspects of new policies, but there is another side of the coin. In the process of conducting the experiment and analyzing the data, the investigators will learn a great deal about "what's going on," and this increased knowledge will ordinarily lead to significant improvements in the programs under consideration, no matter how well-conceived the programs were initially. Purely analytical studies based on nonexperimental data can never provide such realistic lessons. During the course of the experiment, a number of characteristics that may become relevant in evaluating the experiment will be recorded for each experimental subject (recruit, unit, location, and so on), and additional data on the experimental subject can ordinarily be acquired at negligible cost. Other types of information—anecdotal reports from administrators, feedback from the experimental subjects themselves, unexpected departures of the experimental responses from previously accepted norms—will add to the data file. In this way even very simple experiments often yield an abundance of data that may be mined for information having important policy implications. For example, HumRRO's experiments to compare different basic training techniques not only provided comparisons between two training methods in terms of test scores, dropout rates, and time required for mastery of the course material, but also showed that the trainees in mental category IV (those scoring between the 10th and 30th percentiles on the Armed Forces Qualification Test) performed almost as well as those in the other categories.[6]

Controlled experiments to test alternative personnel programs, like clinical trials to test drugs, provide a fertile setting for making discoveries that may lead to favorable returns far beyond the original goals of the experiment. Although the results may not measure up to Columbus' hard-to-finance experiment to find a new route to India, or Sir Alexander Fleming's accidental discovery of penicillin, one can almost count on making some unforeseen discoveries in the course of conducting and analyzing a carefully controlled experiment. The procedure of taking systematic observations, followed by the application of modern statistical techniques that deliberately search for outlying individual observations and unusual patterns among the variables, lends itself to serendipity in the same way that clinical trials often inadvertently turn up remarkable cures.

[6] K. Weingarten, J. Hungerland, and M. Brennan, *Development and Implementation of a Quality-assured, Peer-instructional Model*, Human Resources Research Organization, Technical Report 72-35, Alexandria, Virginia, November 1972, pp. 17-18.

Consider, for example, an experiment to evaluate the effectiveness of certain recruiting incentives in which the incentives are assigned to specially selected units and the other units are used as a control group. In the process of the experiment it will surely be discovered that certain units in the control group seem to be doing unusually well (or unusually poorly) compared with units having similar characteristics. In addition to evaluating the effects of the incentives, the statistician is afforded the opportunity to isolate those units in the control group whose recruiting performances deviate markedly from the pattern of the others. Are the "outliers" attributable to outstanding recruiters, unusually productive recruiting techniques, peculiar economic phenomena, or just unexplainable "randomness"? Questioning the recruiters and the commanders of the outstanding units may lead to a successful national recruiting campaign or to an improved method of screening recruiters. A pattern of poor recruiting performance in college towns or in high-income cities or in sparsely populated states may indicate undesirable locations for new unit assignments.

The HumRRO studies that were undertaken to compare different teaching methods for basic training show how important unexpected findings can be:

> One cadreman was to train two men so that they could pass a strict performance test for phone installation. When they had passed this test, each man would train two more men. If these four subjects could pass the test, they, in turn, would teach the task to eight more trainees, and so on.
>
> The research staff carried out the experiment for four generations, or to the point where 28 trainees had been trained by other subjects. Every trainee, regardless of aptitude, passed the test, and—what is more startling—the time required for a man to learn the task declined from generation to generation. The training of one student by another—or peer instruction—not only was just as effective as cadre instruction, but was also more efficient in terms of time. . . .
>
> The recognition of the full potential of peer instruction required the catalysis of the further clarification of cost constraints and two unexpected findings. . . . Students invariably preferred direct "hands on" experience with actual equipment to learning via mock-ups and other "sophisticated" media. In addition, students reported a great deal of informal, unplanned peer instruction and stressed the value of peer instruction in their learning.[7]

The other benefits of pilot studies result from the fact that, in the process of implementing the program on a small scale, one must necessarily cross most of the administrative hurdles required to implement the program across the board. Perhaps the pilot study itself requires legal changes, reassignment of personnel, modifications of existing performance criteria, new manuals, or new hardware. Not only will the pilot study greatly facilitate the implementation of the servicewide program, but the process of making the change itself will have undergone a test. Moreover, the close monitoring of the experimental responses and the determination of performance criteria for the alternative schemes under test often lead to valuable quality control methods and measures of performance for the system once the new program is under way.

[7] Ibid., p. 5.

V. ALTERNATIVES TO CONTROLLED EXPERIMENTATION

Some "alternatives" to controlled experimentation for gathering information about proposed personnel programs are: (1) expert opinion, (2) anecdotal evidence, (3) sample surveys, (4) analyses based on managerial records and other nonexperimental data, and (5) simulation. As I indicated earlier, comprehensive pilot studies have many goals, some of which could not otherwise be accomplished. Moreover, pilot studies typically incorporate several of the above methods as an integral part of the experimental program, and the alternatives should not necessarily be regarded as competitors. Nevertheless, some comparisons can be made by an examination of the attributes and shortcomings of the other methods, to lay a basis for specifying those circumstances in which pilot studies are preferable.

EXPERT OPINION

Military leaders have often relied heavily on expert opinion in devising and implementing personnel programs. The studies prepared for the Gates Commission provided expert guidance, as well as historical data and analyses, for programs that might be undertaken during the transition to an all-volunteer force.[8] A more commonly used device for generating "expert" opinion is characterized by the interoffice memorandum from the military leader to the members of his staff asking for an in-depth analysis of a program within 24 hours. Expert opinion can be wrong, misleading, and biased by selfish motives or group pressure; one would hope that this is the exception not the rule, but the experts may lack the high-quality information necessary for making clear-cut evaluations of complex programs. In the absence of controlled experiments in certain areas, their expertise is limited to educated guesses.

Although I recommend that controlled experiments be used to reduce the extent to which personnel management decisions are based on expert opinion, I recognize that the pilot studies themselves rely heavily on, and profit greatly from, expert opinion. The guidance of those who are most knowledgeable about the problems

[8] *Studies Prepared for the President's Commission on an All-volunteer Armed Force,* Vols. I & II, Washington, D.C., November 1970.

under investigation is essential to all stages of the controlled experiment. Experts are needed to help formulate the program, to specify the criteria by which the program's performance should be gauged, to help design the study, to pinpoint the variables that may have a significant bearing on the process under study, and to bring relevant textbook theory and past research to bear on the conduct and analysis of the experiment.

ANECDOTAL EVIDENCE

Anecdotal evidence consists of informal comments by people who report their observations and criticisms of the system under study. Such evidence might be gathered through a "Suggestion Box" or by soliciting comments at the end of an opinion survey. In medical research the carefully documented cases reported in medical journals are an important type of anecdotal evidence leading to the development of lists of symptoms associated with certain diseases and reports of successful treatments. In the HumRRO studies of different training methods, feedback from the students led to significant improvements in the training program.

Anecdotal evidence plays an important role in pilot studies, but there are dangers inherent in its analysis that detract from its potential usefulness. First of all, data of this type are usually "selected data"; that is, those who volunteer the information are atypical of the population to which the program is directed. Students who drop by after class to provide feedback to their instructors tend to have a far rosier picture of the course than those who don't. In general, this type of evidence, coming from a variety of sources of unknown reliability and with unknown biases, defies ordinary statistical analysis. In those cases where anecdotal data are systematically collected from all participants in the study or from a randomly selected sample, the informal opinions may still be misleading. The people who work within a particular system day after day become stereotyped in their thinking about their activities and fail to consider alternatives that would occur to outsiders. They may dismiss the possibility of improving their system, or they may react defensively to observers, especially when their daily routines are threatened. Yet, the discovery of important innovations often stems from anecdotal evidence, and solicitation of this type of information from the participants in a pilot study may lead to significant improvements in the programs and enhance the discovery of new knowledge.

SAMPLE SURVEYS

Sample surveys are more systematic than anecdotal evidence for gathering opinions from members of the population that might be affected by the proposed programs. Ordinarily, they are conducted by having respondents fill out questionnaires. These respondents have been specially selected using well-established sampling techniques to assure that they are representative of the population under

study. The statistical theory behind sample surveys is well developed,[9] and practice has become more and more consistent with theory. Sample surveys are versatile in that a large number of questions can be asked in an attempt to determine precise measures of the relationships between the inputs and the outputs of the different versions of the program. By soliciting information on the characteristics of the respondents, one can see how various subpopulations might respond to various options, thus permitting the program managers to select those options that would have greatest appeal to certain subpopulations. Moreover, the method is cheap, especially if one can ignore the time spent by the respondents in filling out the questionnaires.

There are some common misconceptions about the shortcomings of sample surveys that must be considered in making comparisons between sample surveys and pilot studies. Since a sample size of only a few thousand seems tiny relative to a population size of millions, many people refuse to accept the fact that sample surveys can provide very precise estimates of population parameters, especially when carefully stratified samples are used. Also, people attach undue precision to "complete" samples in which supposedly every member of the population is questioned. As the Bureau of the Census learned long ago, the complete decennial censuses are subject to gross errors, and the Bureau has been using results of its sample surveys to adjust the census counts for some time. For one thing, the persons who conduct the sample surveys are more experienced and better trained than the survey takers; also, the much smaller sample size enables better data-editing and quality control. Perhaps people attach great reliability to the electronic data-processing machines that handle the complete surveys, but the output from these machines can be no more accurate than the records used as input, and a few records that are flagrantly in error or omit key bits of information offset the precision theoretically achievable through complete counts.

Insofar as evaluating proposed personnel programs is concerned, the major shortcoming of the sample survey is not due to the statistical unreliability of the sampling process but to the factual unreliability of the respondents themselves. No matter how large the sample size, if a small proportion of the respondents deliberately or even unconsciously distort their answers in the same way, the results of the survey are biased. As an example below will indicate, the unknown extent of the bias can negate the information collected. The type of survey that is especially prone to this kind of distortion is one in which the respondent is asked what action he would take if a certain program were initiated. The purpose of such a survey is to pool the individual responses to predict what would happen *if* the program were to be implemented, whereas a pilot study would measure the actual responses of the participants *when* the program is initiated. Such surveys suffer by comparison with pilot studies because of artificiality of the survey technique in the same way that responses to the question "What would you do if I gave you a million dollars?" differ considerably from the actions of individuals who actually receive a million dollars.

The lack of realism associated with filling out a questionnaire in a survey of this type leads to a certain amount of "gamesmanship" on the part of the respondents.

[9] See W. Edward Deming and Alan Stuart, "Sample Surveys," *International Encyclopedia of the Social Sciences*, Vol. 13, The Macmillan Company & The Free Press, Riverside, New Jersey, 1968, pp. 596-616, and the references cited there.

No matter what degree of anonymity is assured the respondents, some will distort their answers to conform to what they feel are "normal" or desirable responses. Others will try to answer in such a way as to bring about changes that they feel would benefit themselves or their friends. Still others may attempt to confound the statisticians. Even if the respondent is trying to answer the questions conscientiously, his quick answers to a large number of questions may not correspond to his responses after giving the questions more consideration. The questions themselves may not be clear to the respondents, and a trivial rewording of a question or a reordering of a sequence of questions may elicit substantially different results.

The Research Analysis Corporation conducted a sample survey in late 1971 that may have suffered from many of the shortcomings listed above. 4752 enlisted men were drawn from a population of 524,000 reservists in the Army Reserve Forces who were in their sixth (and final) year of their initial enlistment tours.[10] Fewer than 50 percent of the men in the survey responded to the mailed questionnaire, which in itself may have prejudiced the findings.

To determine the potential effectiveness of various economic incentives in raising reenlistment rates in the Army Reserve Forces, the reservists were asked ten questions related to ten different incentives. One of the questions read as follows:

> Suppose you were allowed as a member of the Reserve Components to carry $15,000 of Servicemen's Group Life Insurance (SGLI) at a nominal cost to you of, say, $3 to $4 per month. *Considering only this item,* what is the probability that you would have enlisted in the Guard or Reserve if there were no military draft.

In responding, each reservist was to encircle one of the numbers 0, 10, 20, . . . , 100 indicating his "percent probability of enlisting." After averaging the responses to these questions, RAC listed the "mean probability of reenlistment" for each of the incentives. They are listed below for the Army National Guard *in the same order as the questions were asked:*

No additional incentives	19%
Increased pay	43%
Bonus ($100 per year)	32%
No loss of income during training	32%
Education benefits ($500/3 yrs.)	32%
SGLI at nominal cost	27%
Improved retirement benefits	41%
Medical and dental benefits for reservist and his family	50%
Home loan guaranteed by FHA or VA	45%
Proficiency pay	34%
Federal income tax exemption ($1250)	48%

First of all, as the question quoted above indicates, the questions did not specifically ask for the reservists' subjective estimates of their reenlistment probabilities, but RAC analysts apparently drew that inference from the survey. At any rate, the increases in the reenlistment rates attributable to these incentives, as suggested by the list, are surely inflated because of the factors mentioned above. Note that the three supposedly most effective incentives correspond to three out of the last four

[10] W. L. Clement et al., *Maintenance of Reserve Components in a Volunteer Environment,* Research Analysis Corporation, RAC R-148, McLean, Virginia, November 1972.

questions on the list. Although the entire sentence beginning "*Considering only this item,*" was repeated with each question, some reservists may have ignored the phrase after reading it once or twice. If so, the incentives near the bottom of the list may not be as appealing as those near the top, despite their higher percentages. Since the extent of the overstatement of the percentages because of gamesmanship, misunderstanding of the questions, overly quick responses without due consideration of the true worth of the incentives, and "nonrespondent bias" cannot be estimated, the value of the survey to decisionmakers is greatly reduced. Although one can see ways to improve this particular survey, surveys of this type can never be more reliable than the respondents; and man is often an imperfect observer and fanciful reporter of his own behavior.

ANALYSIS OF NONEXPERIMENTAL DATA

Analysis of managerial records, time series data, and other nonexperimental data provides another alternative to experimentation. Associated with almost every personnel process are people who systematically collect records, measures of performance, and other data that can be exploited to yield information about existing personnel programs. Observational studies based on historical data are also widely used to estimate the potential effects of introducing new personnel programs. Usually the rationale behind these studies can be stated roughly as follows: *The measures of output (the dependent variables) were approximately related to the values of other (independent) variables according to certain formulas. If the new policy consists of changing one or more of the independent variables and the values of the other independent variables either remain unchanged or change in a deterministic way, then the outputs under the new policy can be estimated by substituting the new values of the independent variables in the formulas.* Clearly, this rationale can be challenged on so many grounds that evaluations of new programs obtained in this way are highly suspect, however careful the study. In brief, such studies are subject to the following hazards:

1. The historical data may be inaccurate, incomplete, or even deliberately contrived as a result of personnel turnover, inconsistent practices over time, changing definitions, or "fudging" of the data to make the program look good.
2. The evaluators are tied to available data bases, so that the evaluations must either exclude some important factors for which no measurements are available or substitute surrogate measures of questionable reliability.
3. The formulas derived from the historical data may be distorted because of the incorrect specifications of the functional form of the prediction formulas, the exclusion of important missing variables, or biases in the data.
4. Even if the formulas describe the historical data well, they may be worthless as predictors of results attainable under the new policy.

It is implicitly assumed in using formulas that have been fitted to observational data that the formulas describe a cause-and-effect relationship between the independent variables and the measures of output. However, it is usually difficult, if

not impossible, to ascertain to what extent the magnitude of the relationship results from causal links among the variables and to what extent it reflects the way the variables are associated with each other because of their mutual dependence on latent variables, errors of measurement, and random fluctuations.[11] Changing the levels of the independent variables to achieve the results promised by formulas amounts to wishful thinking. If the predictions involve extrapolations beyond the range of the historical data, further hazards come into play, partly for statistical reasons, but also because a formula may provide an excellent approximation of a relationship among variables over one region but fit very poorly over another region.

Some of the hazards mentioned above are absent in controlled experiments. Also, the effects of the disturbing variables that mask the true cause-and-effect relationships in observational studies can be controlled in pilot studies by (1) selecting the experimental subjects in such a way as to hold certain factors constant; (2) choosing the experimental environment to minimize the effects of extraneous factors; (3) using statistical techniques (randomization, blocking, replication, analysis of covariance adjustments) either to suppress the effects of both known and latent variables or to facilitate the estimation of the effects of the known variables more precisely. Either case leads to more accurate comparisons of the options under test.

To indicate the shortcomings of observational studies relative to pilot studies, compare how the two approaches might have been used to analyze the feasibility of relaxing mental qualification standards for enlisted men. This proposal was actually tested through a well-conceived pilot study called "Project 100,000" beginning in October 1966.[12] Over a three-year period, the military services relaxed their mental and physical standards, accepting 246,000 "New Standards" men who would otherwise have been disqualified either because of failure to pass the mental tests or because of easily correctable physical defects. These men had percentile scores of between 10 and 20 on the Armed Forces Qualification Test and a median score of 13.6. The overall performance of these men in the service was gauged relative to a control group consisting of the other men who entered the service during the same period. A number of criteria were used, including attrition rates during training, promotion progress, supervisory evaluations, disciplinary records, and reenlistment rates. Since most of the men performed as fully satisfactory servicemen, the experiment not only proved that the reduction in mental standards was feasible but also provided information of continuing value to indicate how men with these attributes might be better used in the future.

Contrast this exemplary pilot study with a hypothetical observational study that might have taken place before Project 100,000. Such a study might have stated, in part: "Based on our analysis of the existing data on attrition rates in certain occupational training programs, we estimate that approximately 13 percent of the New Standards men will become dropouts, as compared with only 4 percent for the men who meet the usual requirements." A little thought about how our hypothetical analyst might arrive at the precise-sounding 13 percent suggests that this number must result from extrapolating a formula fitted to the observed data. However, as

[11] For a theoretical discussion of the hazards of inferring causality in regression situations, see footnote 4, above; and W. G. Cochran, "The Planning of Observational Studies of Human Populations." *The Journal of the Royal Statistical Society*, Series A, Vol. 128, 1965, pp. 234-265.

[12] Assistant Secretary of Defense (M&RA), *Project One Hundred Thousand*, Department of Defense, Washington, D.C., December 1969.

in most cases, the extrapolated value is highly sensitive to the specification of the functional form of the regression equation.[13] Also, this is a situation in which the new policy would surely destroy the historical pattern relating the key variables (attrition rates and AFQT scores). With this much of a change in the mental caliber of the trainees, it seems only reasonable to expect that training techniques would be modified and performance standards would be bent to accommodate the new group.[14] Hence, any attempt to evaluate this particular program by an observational program would be subject to most of the hazards listed above, whereas Project 100,000 lived up to its great potential as a social experiment and source of information.

Despite the shortcomings of observational studies relative to pilot studies in evaluating personnel policies, observational studies will continue to play an essential role in the field of human resources. There are certain types of human experiments that cannot be conducted, and others are not feasible for a variety of reasons. Also, observational studies are ordinarily much less costly than pilot studies, they can be done more quickly, and they provide some of the same benefits. In general, observational studies in the civilian sector are hampered by the lack of high-quality comprehensive data on the individuals under study, but the military services have a rich supply of well-maintained personnel data. That data base should be exploited to a far greater extent than it is now to test the desirability of military personnel policies.

SIMULATION

Simulation refers to a technique for analyzing the behavior of a system through a mathematical model that mimics the operation of the system. Ordinarily the model consists of a set of equations that relate the variables of the system and parameters of interest in such a way that, given the current values of the variables and certain assumptions about future inputs into the system, one can simulate the future progress of the system to analyze its sensitivity to changes in the parameters or in the equations themselves. For example, a model might be used to simulate the flow of personnel through a reserve component so as to analyze changes over time in strength by grade or occupational specialty. The equations might express relationships between numbers of personnel in various categories during a particular time period with those during a previous time period. Given the numbers of personnel in the various categories at a particular time, along with information about the

[13] The 13 percent value results from choosing a regression function of the form $y = 1/(a + bt)$ where t = test score and y = attrition rate. To make the regression function pass through 1.0 when $t = 0$ requires setting $a = 1.0$. The value of b was chosen to give an attrition rate of 4 percent at $t = 50$, which appears consistent with data before Project 100,000. This particular regression function is barely distinguishable from a straight line having negative slope over the interval from $t = 20$ to $t = 100$. Choices of a less than 1.0 or other forms of the regression function, such as $y = a + bt$ or $y = 1/[1 + \exp(a + bt)]$, yield widely different results at $t = 13.6$, the median score for the New Standards men, although these regression functions would all fit the data reasonably well. The actual attrition rate among the New Standards men was approximately 8.5 percent.

[14] To quote the Ginzberg Report, p. 52: "One of the by-products of Project 100,000 was the removal of algebra and physics from the curriculum when it was found these subjects were not necessary for the student's eventual mastery of the skill required to operate standard electronic devices."

numbers of new accessions to be expected during each later time period, one can then simulate the process to derive the component's personnel posture at any later time period.[15] In the simulation process, some of the parameters (for example, the reenlistment rates within various categories) may be varied over a range of values to determine the extent to which changes in them would affect shortages in various categories.

Since such models can be used to forecast the effects of new personnel programs, they become competitors to pilot studies. For example, in the model described above, if one can predict that the enactment of a reenlistment bonus for reservists would result in a certain increase in the reenlistment rates, then the simulation of the model will provide forecasts of the resulting changes in the future personnel posture to be achieved from the bonus. This example begins to suggest some of the shortcomings of simulation. The parameters and equations relating the variables of the process are not spontaneously generated within the computer nor do the blinking lights and revolving disks attract divine intervention. The model and the parameters are human creations built upon subject matter theory, expert opinion, observational studies, anecdotal data, and often just wild guesses. Forecasts from the models may or may not be good ones, depending on the degree of correspondence between the model and the actual system. Measures of the degree of reliability of the forecasts are rarely available. Perhaps it is unrealistic to expect a model to forecast the extent to which a reenlistment bonus will raise reenlistment rates, but it may well be within the realm of current capabilities to use simulation models to indicate what changes would result *if* the reenlistment rates should be raised by a certain percentage, and this would be valuable information in itself.

The simulation of classification strategies illustrates another facet of this technique.[16] A classification strategy is a rule for assigning recruits to occupational specialties based on the recruits' personal characteristics. Suppose that for each strategy under consideration there are several measures of the goodness of the classification for each recruit based on his personal characteristics and the occupational specialty to which he is assigned. Summing the individual measures over all recruits provides overall measures of goodness for each strategy, thereby permitting comparisons among the various strategies on a number of criteria. The characteristics of the recruits may be drawn from past personnel records, or they may be randomly generated using Monte Carlo techniques to be representative of a certain population from which future recruits will be drawn. Using randomly generated data to capture the essence of naturally occurring randomness highlights another common aspect of the simulation process and indicates why "simulation" is ordinarily used in the context of "computer simulation."

Although the computer's role in the above sample is restricted to carrying out large numbers of arithmetic calculations, its capability for performing many operations in a split second permits the modeling of far more complicated systems than in the past. Presumably, the computer's ability to retain more of the complexity of

[15] For an example of a model of this type, see B. Rostker, *Air Reserve Forces Personnel Study: Volume I. The Personnel Structure and Posture of the Air National Guard and the Air Force Reserve*, R-1049-PR, The Rand Corporation, Santa Monica, April 1973.

[16] For a comparison of classification strategies along the lines outlined here, see Edward F. Alf and John H. Wolfe, "Comparison of Classification Strategies by Computer Simulation Methods," U.S. Naval Personnel Activity, Technical Bulletin STB 68-11, San Diego, June 1968.

the original system leads to a better correspondence between the model and the real world. Also, the computer's capability for cheap storage and retrieval of the huge data banks available in the military, such as the longitudinal personnel records of all personnel in a military service, makes it a valuable research tool that has only begun to be exploited.

At the present time, computer simulations are poor substitutes for pilot studies in evaluating new personnel programs. However, there is a clear need for undertaking both types of programs immediately. We should look to the future when more and more reliance will be placed on computer models to simulate personnel processes. In the meantime, we urgently need some pilot studies to evaluate proposals now under consideration, and these in turn will provide some of the hard facts needed for constructing the computer models of the future.

VI. SOME SHORTCOMINGS OF PILOT STUDIES

After praising the glories of pilot studies and demeaning the alternatives, I now turn to some of the difficulties inherent in using controlled experimentation in the military. In brief, the problems are these:

1. If the "treatments" involve providing special privileges to or depriving selected personnel, questions of inequity may arise. From my experiences in designing experiments for the Air Force, military leaders seem particularly sensitive to this aspect of experimentation. However, as in medical trials to test new drugs or different treatments, possible inequalities associated with using different treatments for different groups of people must be weighed against the benefits. Given the amount of money and time involved and the alternative uses that could be made of savings achieved through finding more efficient personnel programs, I certainly cannot place the "unfairness" associated with granting special privileges to a tiny proportion of the men on the same scale as the ethical obligation to the public to test some of these programs before they are implemented.

2. Pilot studies can be very expensive. However, many of the costs attendant on the implementation of the pilot study are start-up costs that the new program, if adopted, would have to bear anyway. Moreover, the pilot study may lead to a reduction in the costs of the full implementation through efficiencies learned in the pilot study.

3. Pilot studies are time-consuming. Designing the experiment, convincing everyone concerned that the test is worthwhile, and making the changes required to implement the test may take a substantial amount of time. Once the test is under way, an additional period may be required to get the "bugs" out and let the newness wear off. It may take several months before the true effects of the programs can be properly measured, and by the time the data are analyzed and the reports are written, the entire study may take a year or more.

4. The experimental nature of the pilot study may affect the findings. The responses of the participants may be inflated by "Hawthorne effects." That is, if the experimental subjects like the idea of participating in the experiment or they anticipate that their behavior will be monitored closely, they may respond unusually well. This effect, which is analogous to the "placebo effect" in medicine, is a common phenomenon in social experiments.[17] Also, if the experimental situation seems too

[17] For a discussion of Hawthorne effects in various situations, see Robert Rosenthal and Lenore Jacobson, *Pygmalion in the Classroom*, Holt, Rinehart and Winston, New York, 1968.

artificial, the responses may be biased by the same factors that were alluded to in the discussion of sample surveys.

5. Few large-scale pilot studies have taken place in the military. Extra time and effort may be required to put the study into effect because of the natural reluctance on the part of military leaders to undertake the risks that they see in using pilot studies to evaluate programs. There are few precedents, and little experience has been accumulated in carrying them out. Since the fields that have benefited most from controlled experiments—agriculture, medicine, and the physical sciences—are remote from the military, neither the benefits nor the advanced state of the field of experimental design are well known.

Expending a lot of time and money to test promising programs runs counter to standard operating procedure in the military. If a policymaker faces a critical problem and some immediate action promises to alleviate that problem, the natural reaction is to implement the policy as soon as possible and forget the test, even though this kneejerk response in the past has contributed to the current predicament. Since pilot studies of important programs are not risk-free enterprises, it will require enlightened leadership on the part of some to reverse this pattern.

VII. THE EXPERIMENTAL DESIGN

Many statistical and operational considerations are involved in designing a controlled experiment. Although a number of excellent books have been written on the subject,[18] the peculiarities of each pilot study are ordinarily such that general theory provides only partial guidance.

The experiment should be designed to enhance the validity and precision of the analysis to the greatest extent possible subject to the operational and cost constraints that are imposed. In general, the type of analysis used to compare (and estimate the interactions of) the various treatments under study ordinarily falls within the framework of either "analysis of variance" or "analysis of covariance" (regression using both quantitative and qualitative variables).[19] In such cases, the experimental design amounts to choosing the experimental subjects and determining which subjects and how many should be assigned to each treatment group. The assignment depends on a variety of factors including operational constraints, reliability of estimates, experimental and analytical costs, and administrative convenience. The sizes of the treatment groups may be dictated by operational necessity or convenience, but more likely they will be chosen through a consideration of the tradeoff between increased precision through larger sample sizes and costs of carrying out the experiment.

Usually the experimental subjects or units will have certain distinguishing characteristics, such as the values of background variables for trainees. These characteristics may be used in the analysis in making comparisons between the treatment and control groups. For statistical reasons, the distributions of the characteristics of the experimental units should be similar over all the treatment and control groups. If the sample sizes are large, this similarity can be achieved by assigning the experimental units to the treatment and control groups at random. The results of the randomization can then be checked to ascertain how closely the distributions match. In the event that the randomization should lead to a marked imbalance in the distributions of these characteristics, a second randomization can be performed. The advantage of randomization is that latent variables that cannot be measured or are unknown to the investigators are also likely to be controlled for by the randomization process. These statistical techniques reduce the biases in making

[18] See, for example, D. R. Cox, *Planning of Experiments,* John Wiley & Sons, Inc., New York, 1958.

[19] These statistical techniques are discussed in depth in Henry Scheffé, *The Analysis of Variance,* John Wiley & Sons, Inc., New York, 1959.

comparisons between the treatment and control groups, increase the reliability of the comparisons, increase the precision of other estimates, and enhance the process of searching for outliers and unusual patterns in the data that may lead to increased knowledge about the process and the subjects under investigation.

In some studies, the number of experimental units available for experimentation is quite small. In this case, one cannot rely on unconstrained randomization to achieve the balancing among the groups that is desirable for statistical purposes. This situation, which occurred in a pilot study to test the feasibility of reducing the enlistment term in the Air Force Reserves, will be discussed in detail in another paper.

Sometimes opportunities to conduct pilot studies occur naturally in that a new scarce resource becomes available, and one can take advantage of the scarcity by assigning the resource to those experimental units randomly (or deliberately, if the sample size is small) so that the units chosen to receive the resource will be comparable to those that were not. For example, the Air Reserve Forces recently received funding to assign recruiters to their reserve units, but there was a period during which recruiters were available to man only some of the bases. This set up a situation for comparing the recruiting performances of those units that had recruiters with those that had none, providing a pilot study to analyze the effectiveness of the recruiters. The methodology and the results of this experiment will be published in a subsequent study.

Since a number of considerations of a technical nature are involved in setting up a pilot study, it is advisable to have the services of a statistician who is particularly knowledgeable in the field of experimental design. He must necessarily work closely with people who know the inner workings of the process under study. An analysis of the challenges made against recent large-scale pilot studies in the civilian sector reveals the necessity for incorporating expert opinion into the design of the experiment.[20]

[20] For a critical analysis of the Head Start Program, see Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," *Harvard Educational Review,* Vol. 40, Winter 1970, pp. 51-104. For a rejoinder to Smith and Bissell, see Victor G. Cicirelli, John W. Evans, and Jeffry S. Schiller, "The Impact of Head Start: A Reply to the Report Analysis," *Harvard Educational Review,* Vol. 40, Winter 1970, pp. 105-129.

VIII. THE OBLIGATION TO EXPERIMENT

The military services are involved in distributing scarce resources—manpower, talent, time, equipment, and a tremendous amount of money. Given the magnitude of the costs associated with erroneous decisions, as well as the forgone opportunities to invest these resources in other ways, the military has an obligation to evaluate new programs carefully *before* they are implemented to see how to strengthen them, to determine whether to keep them or scrap them, and to learn from them.

The military can no longer afford exorbitant recruiting and retention strategies. If total budgets continue to be relatively fixed in constant dollars as they have been during the past five years, further increases in personnel costs will necessarily lead to diminished weapon systems acquisition, lower troop strengths, and an overall reduction in defense capability. Thus, efficiencies in personnel management will play a more central role in the mission of the armed forces. Also, now that military pay is more closely comparable to civilian pay, Congress will view further direct or indirect increments in the military pay package with skepticism. From the overall military viewpoint, this implies that the increments chosen must be those that will yield the best results for each additional dollar expended. Therefore, it is crucial that military policymakers have better information on cost-effectiveness options before they are adopted.

Another important reason for undertaking controlled experimentation is that policymakers need to have evaluations of the effects of past policy changes to accumulate experience for making decisions in the future and to incorporate quality control into the decisionmaking process. Some would argue that one can evaluate the effectiveness of a new program after it is implemented across the board by making a simple comparison of relevant statistics observed both before and after the policy went into effect. In fact, it is often extremely difficult, if not totally impossible, to differentiate the effects of a new policy from the effects of other events occurring at the same time.[21] Also, the true response to the new policy may precede or follow the implementation of the policy by a considerable length of time depending on the type of program, the difficulties associated with starting up the program, and the unpredictable effects associated with novel undertakings. For example, the effectiveness of the large pay raise of November 1971 upon first enlistments cannot be determined by comparing the numbers of accessions before and after the raise. Such

[21] Donald T. Campbell, "Measuring the Effects of Social Innovations by Means of Time Series," in Judith M. Tanur, ed., *Statistics: A Guide to the Unknown*, Holden-Day, Inc., San Francisco, 1972.

a comparison, without allowing for increased enlistments in anticipation of the pay raise, reductions in draft calls, seasonal fluctuations, economic factors, the changes in the Southeast Asia situation, and concurrent trends in college attendance rates, might even suggest that the pay raise had a negative effect. Although across-the-board pay raises are not the kinds of policies that can be tested in a pilot study, this example illustrates the complexity and uncertainty associated with doing time series analysis to assess effects of policy changes. The use of a control group in situations like this facilitates the evaluation since the external factors unrelated to the policies under test affect the control and treatment groups in the same way.

One reason that more pilot studies have not been conducted in the past may be that military leaders tend to be action oriented. Doing an experiment means putting off making a decision until more information becomes available and may be labeled as indecisiveness, not a highly valued trait in the military. Furthermore, the policy under consideration probably has its proponents who will argue persuasively using indirect evidence for its immediate adoption. Experimentation admits the possibility that hard data will prove them wrong.

Without experimentation, we are dependent on questionable information to make piecemeal adjustments that are likely to lead to gross inefficiencies. In responding to current exigencies, we implement one policy after another, learning little and kidding ourselves that we did the best possible job at each stage by taking the action that looked best at the time. Never getting proper evaluations of our previous "successes" and never determining which features of the program are effective and which are not, we are always left groping for *the* solution in the next policy change.

As we move to an all-volunteer force, the procurement problems are the ones that seem most urgent at the present time. The time is ripe to make a survey to determine the present state of knowledge of the recruiting process, including prior studies or recruiting efforts in all the services. Despite their shortcomings, time study analyses of the effectiveness of past changes in the pay package, recruiting efforts, and incentives should be undertaken. Then a carefully planned pilot study to test the potential effectiveness of the incentives now under consideration and of various recruiting efforts should be undertaken across the services and the reserve under the coordination of the Department of Defense. Within a year or two we would then have reliable answers to many of the recruiting problems now under study in the military, and we would learn a lot about personnel management in the meantime.